Dimension-free tail inequalities for sums of random matrices

Daniel Hsu[1,2], Sham M. Kakade[2], and Tong Zhang[1]

[1]Department of Statistics, Rutgers University
[2]Department of Statistics, Wharton School, University of Pennsylvania

January 19, 2013

Abstract

We derive exponential tail inequalities for sums of random matrices with no dependence on the explicit matrix dimensions. These are similar to the matrix versions of the Chernoff bound and Bernstein inequality except with the explicit matrix dimensions replaced by a trace quantity that can be small even when the dimension is large or infinite. Some applications to principal component analysis and approximate matrix multiplication are given to illustrate the utility of the new bounds.

1 Introduction

Sums of random matrices arise in many statistical and probabilistic applications, and hence their concentration behavior is of significant interest. Surprisingly, the classical exponential moment method used to derive tail inequalities for scalar random variables carries over to the matrix setting when augmented with certain matrix trace inequalities. This fact was first discovered by Ahlswede and Winter (2002), who proved a matrix version of the Chernoff bound using the Golden-Thompson inequality (Golden, 1965; Thompson, 1965): $\operatorname{tr}\exp(A+B) \leq \operatorname{tr}(\exp(A)\exp(B))$ for all symmetric matrices A and B. Later, it was demonstrated that the same technique could be adapted to yield analogues of other tail bounds such as Bernstein's inequality (Gross et al., 2010; Recht, 2009; Gross, 2009; Oliveira, 2010a,b). Recently, a theorem due to Lieb (1973) was identified by Tropp (2011a,b) to yield sharper versions of this general class of tail bounds. Altogether, these results have proved invaluable in constructing and simplifying many probabilistic arguments concerning sums of random matrices.

One deficiency of these previous inequalities is their explicit dependence on the dimension, which prevents their application to infinite dimensional spaces that arise in a variety of data analysis tasks (*e.g.*, Schölkopf et al., 1999; Rasmussen and Williams, 2006; Fukumizu et al., 2007; Bach, 2008). In this work, we prove analogous results where the dimension is replaced with a trace quantity that can be small even when the dimension is large or infinite. For instance, in our matrix generalization of Bernstein's inequality, the (normalized) trace of the second moment matrix appears instead of the matrix dimension. Such trace quantities can often be regarded as an intrinsic

E-mail: `djhsu@rci.rutgers.edu`, `skakade@wharton.upenn.edu`, `tzhang@stat.rutgers.edu`

notion of dimension. The price for this improvement is that the more typical exponential tail e^{-t} is replaced with a slightly weaker tail $t(e^t - t - 1)^{-1} \approx e^{-t+\log t}$. As t becomes large, the difference becomes negligible. For instance, if $t \geq 2.6$, then $t(e^t - t - 1)^{-1} \leq e^{-t/2}$.

There are some previous works that give dimension-free tail inequalities in some special cases. Rudelson and Vershynin (2007) prove exponential tail inequalities for sums of rank-one matrices by way of a key inequality of Rudelson (1999) (see also Oliveira, 2010a). Magen and Zouzias (2011) prove tail inequalities for sums of low-rank matrices using non-commutative Khintchine moment inequalities, but fall short of giving an exponential tail inequality. In contrast, our results are proved using a natural matrix generalization of the exponential moment method.

2 Preliminaries

Let $\xi_1, \ldots, \xi_n$ be random variables, and for each $i = 1, \ldots, n$, let $X_i := X_i(\xi_1, \ldots, \xi_i)$ be a symmetric matrix-valued functional of $\xi_1, \ldots, \xi_i$. We use $\mathbb{E}_i[\,\cdot\,]$ and shorthand for $\mathbb{E}[\,\cdot \mid \xi_1, \ldots, \xi_{i-1}]$. For any symmetric matrix H, let $\lambda_{\max}(H)$ denote its largest eigenvalue, $\exp(H) := I + \sum_{k=1}^{\infty} H^k/k!$, and $\log(\exp(H)) := H$.

The following convex trace inequality of Lieb (1973) was also used by Tropp (2011a,b).

Theorem 1 (Lieb, 1973)**.** *For any symmetric matrix H, the function $M \mapsto \operatorname{tr}\exp(H + \log(M))$ is concave in M for $M \succ 0$.*

The following lemma due to (Tropp, 2011b) is a matrix generalization of a scalar result due to Freedman (1975) (see also Zhang, 2005), where the key is the invocation of Theorem 1. We give the proof for completeness.

Lemma 1 (Tropp, 2011b)**.** *For any constant symmetric matrix X_0,*

$$\mathbb{E}\left[\operatorname{tr}\exp\left(\sum_{i=0}^{n} X_i - \sum_{i=1}^{n} \ln \mathbb{E}_i\left[\exp(X_i)\right]\right)\right] \leq \operatorname{tr}\exp(X_0). \tag{1}$$

Proof. By induction on n. The claim holds trivially for $n = 0$. Now fix $n \geq 1$, and assume as the inductive hypothesis that (1) holds with n replaced by $n - 1$. In this case,

$$\begin{aligned}
\mathbb{E}\left[\operatorname{tr}\exp\left(\sum_{i=0}^{n} X_i - \sum_{i=1}^{n} \log \mathbb{E}_i\left[\exp(X_i)\right]\right)\right] &= \mathbb{E}\left[\mathbb{E}_n\left[\operatorname{tr}\exp\left(\sum_{i=0}^{n-1} X_i - \sum_{i=1}^{n} \log \mathbb{E}_i\left[\exp(X_i)\right] + \log\exp(X_n)\right)\right]\right] \\
&\leq \mathbb{E}\left[\operatorname{tr}\exp\left(\sum_{i=0}^{n-1} X_i - \sum_{i=1}^{n} \log \mathbb{E}_i\left[\exp(X_i)\right] + \log \mathbb{E}_n\left[\exp(X_n)\right]\right)\right] \\
&= \mathbb{E}\left[\operatorname{tr}\exp\left(\sum_{i=0}^{n-1} X_i - \sum_{i=1}^{n-1} \log \mathbb{E}_i\left[\exp(X_i)\right]\right)\right] \\
&\leq \operatorname{tr}\exp(X_0)
\end{aligned}$$

where the first inequality follows from Theorem 1 and Jensen's inequality, and the second inequality follows from the inductive hypothesis. □

3 Exponential tail inequalities for sums of random matrices

3.1 A generic inequality

We first state a generic inequality based on Lemma 1. This differs from earlier approaches, which instead combine Markov's inequality with a result similar to Lemma 1 (*e.g.*, Tropp, 2011a, Theorem 3.6).

Theorem 2. *For any $\eta \in \mathbb{R}$ and any $t > 0$,*

$$\Pr\left[\lambda_{\max}\left(\eta\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\log\mathbb{E}_i\left[\exp(\eta X_i)\right]\right) > t\right] \leq \operatorname{tr}\left(\mathbb{E}\left[-\eta\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\log\mathbb{E}_i\left[\exp(\eta X_i)\right]\right]\right)\cdot(e^t - t - 1)^{-1}.$$

Proof. Fix a constant matrix X_0, and let $A := \eta\sum_{i=0}^{n} X_i - \sum_{i=1}^{n}\log\mathbb{E}_i[\exp(\eta X_i)]$. Note that $g(x) := e^x - x - 1$ is non-negative for all $x \in \mathbb{R}$ and increasing for $x \geq 0$. Letting $\{\lambda_i(A)\}$ denote the eigenvalues of A, we have

$$\begin{aligned}
\Pr\left[\lambda_{\max}(A) > t\right](e^t - t - 1) &= \mathbb{E}\left[\mathbb{1}\left[\lambda_{\max}(A) > t\right](e^t - t - 1)\right] \\
&\leq \mathbb{E}\left[e^{\lambda_{\max}(A)} - \lambda_{\max}(A) - 1\right] \\
&\leq \mathbb{E}\left[\sum_i \left(e^{\lambda_i(A)} - \lambda_i(A) - 1\right)\right] \\
&= \mathbb{E}\left[\operatorname{tr}(\exp(A) - A - I)\right] \\
&\leq \operatorname{tr}(\exp(X_0) + \mathbb{E}[-A] - I)
\end{aligned}$$

where the last inequality follows from Lemma 1. Now we take $X_0 \to 0$ so $\operatorname{tr}(\exp(X_0) - I) \to 0$. $\square$

3.2 Some specific bounds

We now give some specific bounds as corollaries of Theorem 2. Most of the estimates used in the proofs are taken from previous works (*e.g.*, Ahlswede and Winter, 2002; Tropp, 2011a); the main point here is to show how these previous techniques can be combined with Theorem 2 to yield new tail inequalities with no explicit dependence on the matrix dimension.

First, we give a bound under a subgaussian-type condition on the distribution.

Theorem 3 (Matrix subgaussian bound)**.** *If there exists $\bar{\sigma} > 0$ and $\bar{k} > 0$ such that for all $i = 1, \ldots, n$,*

$$\begin{aligned}
\mathbb{E}_i[X_i] &= 0 \\
\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\log\mathbb{E}_i\left[\exp(\eta X_i)\right]\right) &\leq \frac{\eta^2\bar{\sigma}^2}{2} \\
\mathbb{E}\left[\operatorname{tr}\left(\frac{1}{n}\sum_{i=1}^{n}\log\mathbb{E}_i\left[\exp(\eta X_i)\right]\right)\right] &\leq \frac{\eta^2\bar{\sigma}^2\bar{k}}{2}
\end{aligned}$$

for all $\eta > 0$ almost surely, then for any $t > 0$,

$$\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) > \sqrt{\frac{2\bar{\sigma}^2 t}{n}}\right] \leq \bar{k}\cdot t(e^t - t - 1)^{-1}.$$

Proof. We fix $\eta := \sqrt{2t/(\bar{\sigma}^2 n)}$. By Theorem 2, we obtain

$$\begin{aligned}
\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n\eta}\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right) > \frac{t}{n\eta}\right] &\le \operatorname{tr}\left(\mathbb{E}\left[\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right]\right) \cdot (e^t - t - 1)^{-1} \\
&\le \frac{n\eta^2\bar{\sigma}^2\bar{k}}{2} \cdot (e^t - t - 1)^{-1} \\
&= \bar{k} \cdot t(e^t - t - 1)^{-1}.
\end{aligned}$$

Now suppose

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n\eta}\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right) \le \frac{t}{n\eta}.$$

This implies for every non-zero vector u,

$$\frac{u^\top\left(\frac{1}{n}\sum_{i=1}^n X_i\right)u}{u^\top u} \le \frac{u^\top\left(\frac{1}{n\eta}\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right)u}{u^\top u} + \frac{t}{n\eta} \le \lambda_{\max}\left(\frac{1}{n\eta}\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right) + \frac{t}{n\eta}$$

and therefore

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \le \lambda_{\max}\left(\frac{1}{n\eta}\sum_{i=1}^n \log \mathbb{E}_i\left[\exp(\eta X_i)\right]\right) + \frac{t}{n\eta} \le \frac{\eta\bar{\sigma}^2}{2} + \frac{t}{n\eta} = \sqrt{\frac{2\bar{\sigma}^2 t}{n}}$$

as required. ∎

We can also give a Bernstein-type bound based on moment conditions. For simplicity, we just state the bound in the case that the $\lambda_{\max}(X_i)$ are bounded almost surely.

Theorem 4 (Matrix Bernstein bound)**.** *If there exists $\bar{b} > 0$, $\bar{\sigma} > 0$, and $\bar{k} > 0$ such that for all $i = 1, \ldots, n$,*

$$\begin{aligned}
\mathbb{E}_i[X_i] &= 0 \\
\lambda_{\max}(X_i) &\le \bar{b} \\
\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_i[X_i^2]\right) &\le \bar{\sigma}^2 \\
\mathbb{E}\left[\operatorname{tr}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_i[X_i^2]\right)\right] &\le \bar{\sigma}^2\bar{k}
\end{aligned}$$

almost surely, then for any $t > 0$,

$$\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) > \sqrt{\frac{2\bar{\sigma}^2 t}{n}} + \frac{\bar{b}t}{3n}\right] \le \bar{k} \cdot t(e^t - t - 1)^{-1}.$$

Proof. Let $\eta > 0$. For each $i = 1, \ldots, n$,

$$\exp(\eta X_i) \preceq I + \eta X_i + \frac{e^{\eta\bar{b}} - \eta\bar{b} - 1}{\bar{b}^2} \cdot X_i^2$$

and therefore

$$\log \mathbb{E}_i\big[\exp(\eta X_i)\big] \preceq \frac{e^{\eta\bar{b}} - \eta\bar{b} - 1}{\bar{b}^2} \cdot \mathbb{E}_i\big[X_i^2\big].$$

Since $e^x - x - 1 \le x^2/(2(1-x/3))$ for $0 \le x < 3$, we have by Theorem 2

$$\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) > \frac{\eta\bar{\sigma}^2}{2(1-\eta\bar{b}/3)} + \frac{t}{\eta n}\right] \le \frac{\eta^2\bar{\sigma}^2\bar{k}n}{2(1-\eta\bar{b}/3)} \cdot (e^t - t - 1)^{-1}$$

provided that $\eta < 3/\bar{b}$. Choosing

$$\eta := \frac{3}{\bar{b}} \cdot \left(1 - \frac{\sqrt{2\bar{\sigma}^2 t/n}}{2\bar{b}t/(3n) + \sqrt{2\bar{\sigma}^2 t/n}}\right)$$

gives the desired bound. □

3.3 Discussion

The advantage of our results here over previous exponential tail inequalities for sums of random matrices is the absence of explicit dependence on the matrix dimensions. Indeed, all previous tail inequalities using the exponential moment method (either via the Golden-Thompson inequality or Lieb's trace inequality) are roughly of the form $d \cdot e^{-t}$ when the matrices in the sum are $d \times d$ (Ahlswede and Winter, 2002; Gross et al., 2010; Recht, 2009; Gross, 2009; Tropp, 2011a,b). Our results also improve over the tail inequalities of Rudelson and Vershynin (2007) in that it applies to full-rank matrices, not just rank-one matrices; and also over that of Magen and Zouzias (2011) in that it provides an exponential tail inequality, rather than just a polynomial tail. Thus, our improvements widen the applicability of these inequalities (and the matrix exponential moment method in general); we explore some of these in Subsection 3.4.

One disadvantage of our technique is that in finite dimensional settings, the relevant trace quantity that replaces the dimension may turn out to be of the same order as the dimension d (an example of such a case is discussed next). In such cases, the resulting tail bound from Theorem 4 (say) of $\bar{k} \cdot t(e^t - t - 1)^{-1}$ is looser than the $d \cdot e^{-t}$ tail bound provided by earlier techniques (*e.g.*, Tropp, 2011a).

We note that the matrix exponential moment method used here and in previous work can lead to a significantly suboptimal tail inequality in some cases. This was pointed out by Tropp (2011a, Section 4.6), but we elaborate on it here further. Suppose $x_1, \dots, x_n \in \{\pm 1\}^d$ are i.i.d. random vectors with independent Rademacher entries—each coordinate of x_i is $+1$ or -1 with equal probability. Let $X_i = x_i x_i^\top - I$, so $\mathbb{E}[X_i] = 0$, $\lambda_{\max}(X_i) = \lambda_{\max}(\mathbb{E}[X_i^2]) = d-1$, and $\operatorname{tr}(\mathbb{E}[X_i^2]) = d(d-1)$. In this case, Theorem 4 implies the bound

$$\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top - I\right) > \sqrt{\frac{2(d-1)t}{n}} + \frac{(d-1)t}{3n}\right] \le dt(e^t - t - 1)^{-1}.$$

On the other hand, because the x_i have subgaussian projections, it is known that

$$\Pr\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top - I\right) > 2\sqrt{\frac{71d + 16t}{n}} + \frac{10d + 2t}{n}\right] \le 2e^{-t/2}$$

(Litvak et al., 2005, also see Lemma 2 in Appendix A). First, this latter inequality removes the d factor on the right-hand side. Perhaps more importantly, the deviation term t does not scale with d in this inequality, whereas it does in the former. Thus this latter bound provides a much stronger exponential tail: roughly put, $\Pr[\lambda_{\max}(\sum_{i=1}^n x_i x_i^\top/n - I) > c \cdot (\sqrt{d/n} + d/n) + \tau] \leq \exp(-\Omega(n \min(\tau, \tau^2)))$ for some constant $c > 0$; the probability bound from Theorem 4 is only of the form $\exp(-\Omega((n/d)\min(\tau, \tau^2)))$. The sub-optimality of Theorem 4 is shared by all other existing tail inequalities proved using this exponential moment method. The issue is related to the asymptotic freeness of the random matrices $X_1, \dots, X_n$ (Voiculescu, 1991; Guionnet, 2004)—*i.e.*, that nearly all high-order moments of random matrices vanish asymptotically—which is not exploited in the matrix exponential moment method. This means that the proof technique in the exponential moment method over-counts the contribution of high-order matrix moments that should have vanished. Formalizing this discrepancy would help clarify the limits of this technique, but the task is beyond the scope of this paper. It is also worth mentioning that asymptotic freeness only holds when the X_i have independent entries. For matrices with correlated entries, our bound is close to best possible in the worst case.

3.4 Examples

For a matrix M, let $\|M\|_F$ denote its Frobenius norm, and let $\|M\|_2$ denote its spectral norm. If M is symmetric, then $\|M\|_2 = \max\{\lambda_{\max}(M), \ -\lambda_{\min}(M)\}$, where $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ are, respectively, the largest and smallest eigenvalues of M.

3.4.1 Supremum of a random process

The first example embeds a random process in a diagonal matrix to show that Theorem 3 is tight in certain cases.

Example 1. Let $(Z_1, Z_2, \dots)$ be (possibly dependent) mean-zero subgaussian random variables; *i.e.*, each $\mathbb{E}[Z_i] = 0$, and there exists positive constants $\sigma_1, \sigma_2, \dots$ such that

$$\mathbb{E}[\exp(\eta Z_i)] \leq \exp\left(\frac{\eta^2 \sigma_i^2}{2}\right) \quad \forall \eta \in \mathbb{R}.$$

We further assume that $v := \sup_i\{\sigma_i^2\} < \infty$ and $k := \frac{1}{v}\sum_i \sigma_i^2 < \infty$. Also, for convenience, we assume $\log k \geq 1.3$ (to simplify the tail inequality).

Let $X = \operatorname{diag}(Z_1, Z_2, \dots)$ be the random diagonal matrix with the Z_i on its diagonal. We have $\mathbb{E}[X] = 0$, and

$$\log \mathbb{E}[\exp(\eta X)] \preceq \operatorname{diag}\left(\frac{\eta^2\sigma_1^2}{2}, \frac{\eta^2\sigma_2^2}{2}, \dots\right),$$

so

$$\lambda_{\max}\left(\log \mathbb{E}[\exp(\eta X)]\right) \leq \frac{\eta^2 v}{2} \quad \text{and} \quad \operatorname{tr}\left(\log \mathbb{E}[\exp(\eta X)]\right) \leq \frac{\eta^2 v k}{2}.$$

By Theorem 3, we have

$$\Pr\left[\lambda_{\max}(X) > \sqrt{2vt}\right] \leq kt(e^t - t - 1)^{-1}.$$

Therefore, letting $t := 2(\tau + \log k) > 2.6$ for $\tau > 0$ and interpreting $\lambda_{\max}(X)$ as $\sup_i\{Z_i\}$,

$$\Pr\left[\sup_i\{Z_i\} > 2\sqrt{\sup_i\{\sigma_i^2\}\left(\log\frac{\sum_i \sigma_i^2}{\sup_i\{\sigma_i^2\}} + \tau\right)}\right] \leq e^{-\tau}.$$

Suppose the $Z_i \sim \mathcal{N}(0,1)$ are just N i.i.d. standard Gaussian random variables. Then the above inequality states that the largest of the Z_i is $O(\log N + \tau)$ with probability at least $1 - e^{-\tau}$; this is known to be tight up to constants, so the $\log N$ term cannot generally be removed. This fact has been noted by previous works on matrix tail inequalities (*e.g.*, Tropp, 2011a), which also use this example as an extreme case. We note, however, that these previous works are not applicable to the case of a countably infinite number of mean-zero Gaussian random variables $Z_i \sim \mathcal{N}(0,\sigma_i^2)$ (or more generally, subgaussian random variables), whereas the above inequality can be applied as long as the sum of the σ_i^2 is finite. □

3.4.2 Principal component analysis

Our next two examples uses Theorem 4 to give spectral norm error bounds for estimating the second moment matrix of a random vector from i.i.d. copies. This is relevant in the context of (kernel) principal component analysis of high (or infinite) dimensional data (*e.g.*, Schölkopf et al., 1999).

Example 2. Let $x_1, \ldots, x_n$ be i.i.d. random vectors with $\varSigma := \mathbb{E}[x_i x_i^\top]$, $K := \mathbb{E}[x_i x_i^\top x_i x_i^\top]$, and $\|x_i\|_2 \leq \bar{\ell}$ almost surely for some $\bar{\ell} > 0$. Let $X_i := x_i x_i^\top - \varSigma$ and $\hat{\varSigma}_n := n^{-1}\sum_{i=1}^n x_i x_i^\top$. We have $\lambda_{\max}(X_i) \leq \bar{\ell}^2 - \lambda_{\min}(\varSigma)$. Also, $\lambda_{\max}(n^{-1}\sum_{i=1}^n \mathbb{E}[X_i^2]) = \lambda_{\max}(K - \varSigma^2)$ and $\mathbb{E}[\mathrm{tr}(n^{-1}\sum_{i=1}^n \mathbb{E}[X_i^2])] = \mathrm{tr}(K - \varSigma^2)$. By Theorem 4,

$$\Pr\left[\lambda_{\max}\big(\hat{\varSigma}_n - \varSigma\big) > \sqrt{\frac{2\lambda_{\max}(K - \varSigma^2)t}{n}} + \frac{(\bar{\ell}^2 - \lambda_{\min}(\varSigma))t}{3n}\right] \leq \frac{\mathrm{tr}(K - \varSigma^2)}{\lambda_{\max}(K - \varSigma^2)} \cdot t(e^t - t - 1)^{-1}.$$

Since $\lambda_{\max}(-X_i) \leq \lambda_{\max}(\varSigma)$, we also have

$$\Pr\left[\lambda_{\max}\big(\varSigma - \hat{\varSigma}_n\big) > \sqrt{\frac{2\lambda_{\max}(K - \varSigma^2)t}{n}} + \frac{\lambda_{\max}(\varSigma)t}{3n}\right] \leq \frac{\mathrm{tr}(K - \varSigma^2)}{\lambda_{\max}(K - \varSigma^2)} \cdot t(e^t - t - 1)^{-1}.$$

Therefore

$$\Pr\left[\big\|\hat{\varSigma}_n - \varSigma\big\|_2 > \sqrt{\frac{2\lambda_{\max}(K - \varSigma^2)t}{n}} + \frac{\max\{\bar{\ell}^2 - \lambda_{\min}(\varSigma),\ \lambda_{\max}(\varSigma)\}t}{3n}\right] \leq \frac{\mathrm{tr}(K - \varSigma^2)}{\lambda_{\max}(K - \varSigma^2)} \cdot 2t(e^t - t - 1)^{-1}.$$

A similar result was given by Zwald and Blanchard (2006, Lemma 1) but for Frobenius norm error rather than spectral norm error. This is generally incomparable to our result, although spectral norm error may be more appropriate in cases where the spectrum is slow to decay. □

We now show that combining the bound from the previous example with sharper dimension-dependent tail inequalities can sometimes lead to stronger results.

Example 3. Let $x_1, \ldots, x_n$ be i.i.d. random vectors with $\varSigma := \mathbb{E}[x_i x_i^\top]$; let $X_i := x_i x_i^\top - \varSigma$ and $\hat{\varSigma}_n := n^{-1}\sum_{i=1}^n x_i x_i^\top$. For any positive integer $d \leq \mathrm{rank}(\varSigma)$, let $\Pi_{d,0}$ be the orthogonal projector to the d-dimensional eigenspace of $\varSigma$ corresponding to its d largest eigenvalues, and let $\Pi_{d,1} := I - \Pi_{d,0}$. We have

$$\begin{aligned}
\big\|\hat{\varSigma}_n - \varSigma\big\|_2 &\leq \big\|\Pi_{d,0}(\hat{\varSigma}_n - \varSigma)\Pi_{d,0}\|_2 + 2\big\|\Pi_{d,0}(\hat{\varSigma}_n - \varSigma)\Pi_{d,1}\|_2 + \big\|\Pi_{d,1}(\hat{\varSigma}_n - \varSigma)\Pi_{d,1}\|_2 \\
&\leq 2\big\|\Pi_{d,0}(\hat{\varSigma}_n - \varSigma)\Pi_{d,0}\|_2 + 2\big\|\Pi_{d,1}(\hat{\varSigma}_n - \varSigma)\Pi_{d,1}\|_2.
\end{aligned}$$

We can use the tail inequalities from this work to control $\|\Pi_{d,1}(\hat{\Sigma}_n - \Sigma)\Pi_{d,1}\|_2$, and use potentially sharper dimension-dependent inequalities to control $\|\Pi_{d,0}(\hat{\Sigma}_n - \Sigma)\Pi_{d,0}\|_2$.

Let $\Sigma_{d,0} := \Pi_{d,0}\Sigma\Pi_{d,0}$, $\Sigma_{d,1} := \Pi_{d,1}\Sigma\Pi_{d,1}$, $K_{d,1} := \mathbb{E}[(\Pi_{d,1}x_i x_i^\top \Pi_{d,1})^2]$, and assume $\|\Pi_{d,1}x_i\|_2 \leq \bar{\ell}_{d,1}$ for all $i = 1, \ldots, n$ almost surely. Furthermore, suppose there exists $\gamma_{d,0} > 0$ such that for all $i = 1, \ldots, n$ and all vectors α,

$$\mathbb{E}\left[\exp\left(\alpha^\top \Sigma_{d,0}^{-1/2} x_i\right)\right] \leq \exp\left(\gamma_{d,0}\|\alpha\|_2^2/2\right)$$

where $\Sigma_{d,0}^{-1/2}$ is the matrix square-root of the Moore-Penrose pseudoinverse of $\Sigma_{d,0}$. This condition states that every projection of $\Sigma_{d,0}^{-1/2} x_i$ has subgaussian tails. In this case, the tail behavior of $\|\Pi_{d,0}(\hat{\Sigma}_n - \Sigma)\Pi_{d,0}\|_2$ should not depend on the dimensionality d. Indeed, a covering number argument gives

$$\Pr\left[\left\|\Pi_{d,0}\left(\hat{\Sigma}_n - \Sigma\right)\Pi_{d,0}\right\|_2 > 2\gamma_{d,0}\|\Sigma\|_2\left(\sqrt{\frac{71d+16t}{n}} + \frac{5d+t}{n}\right)\right] \leq 2e^{-t/2}$$

for any $t > 0$ (see Lemma 2 in Appendix A). Combining this with the tail inequality from Example 2, we have (for $t \geq 2.6$)

$$\Pr\Bigg[\left\|\hat{\Sigma}_n - \Sigma\right\|_2 > 4\gamma_{d,0}\|\Sigma\|_2\left(\sqrt{\frac{71d+16t}{n}} + \frac{5d+t}{n}\right) + 2\sqrt{\frac{2\lambda_{\max}(K_{d,1} - \Sigma_{d,1}^2)\left(\log\left(\frac{\operatorname{tr}(K_{d,1}-\Sigma_{d,1}^2)}{\lambda_{\max}(K_{d,1}-\Sigma_{d,1}^2)}\right) + t\right)}{n}} + \frac{2\max\{\bar{\ell}_{d,1}^2 - \lambda_{\min}(\Sigma_{d,1}),\ \lambda_{\max}(\Sigma_{d,1})\}\left(\log\left(\frac{\operatorname{tr}(K_{d,1}-\Sigma_{d,1}^2)}{\lambda_{\max}(K_{d,1}-\Sigma_{d,1}^2)}\right) + t\right)}{3n}\Bigg] \leq 4e^{-t/2}. \quad (2)$$

□

Comparisons. We consider the following stylized scenario to compare the bounds from Example 2 and Example 3.

1. The largest d eigenvalues of Σ are all equal to $\|\Sigma\|_2$, and the remaining eigenvalues are smaller and rapidly decaying so $\operatorname{tr}(\Sigma_{d,1})/\|\Sigma\|_2$ is small.

2. $\bar{\ell}^2$ and $\bar{\ell}_{d,1}^2$ are within constant factors of $\operatorname{tr}(\Sigma)$ and $\operatorname{tr}(\Sigma_{d,1})$, respectively; this simply requires that the squared length of any x_i never be more than a constant factor times its expected squared length.

3. $\lambda_{\max}(K - \Sigma^2)$ and $\lambda_{\max}(K_{d,1} - \Sigma_{d,1}^2)$ are within constant factors of $\lambda_{\max}(\Sigma)^2$ and $\lambda_{\max}(\Sigma_{d,1})^2$, respectively; this is similar to the previous condition.

We will also ignore constant and logarithmic factors, as well as the $\gamma_{d,0}$ factors. The bound on $\|\hat{\Sigma}_n\|_2$ from Example 3 then becomes (roughly)

$$\|\Sigma\|_2\left(1 + \sqrt{\frac{d}{n}}\right) + \|\Sigma\|_2\left(\sqrt{\frac{t}{n}} + \frac{t}{n} + \frac{(\operatorname{tr}(\Sigma_{d,1})/\|\Sigma\|_2)t}{n}\right) \quad (3)$$

whereas the bound from Example 2 is

$$\|\Sigma\|_2 + \|\Sigma\|_2 \left(\sqrt{\frac{t}{n}} + \frac{\Big(d + (\mathrm{tr}(\Sigma_{d,1})/\|\Sigma\|_2)\Big)t}{n} \right). \tag{4}$$

The main difference between these bounds is that the deviation term t does not scale with d in (3), but it does in (4), so the exponential tail in the latter is much weaker, as discussed in Subsection 3.3.

We can also compare the bound from Example 3 to the case where the x_i are i.i.d. Gaussian random vectors with mean zero and covariance Σ. Arrange the x_i as columns in a matrix $\hat{A}_n = [x_1|\cdots|x_n]$, so

$$\|\hat{\Sigma}_n\|_2 = \frac{1}{n}\|\hat{A}_n\hat{A}_n^\top\|_2 = \frac{1}{n}\|\hat{A}_n\|_2^2.$$

Note that $\hat{A}_n$ has the same distribution as $\Sigma^{1/2}Z$, where Z is a matrix of independent standard Gaussian random variables. The function $Z \mapsto \|\Sigma^{1/2}Z\|_2 = \|\hat{A}_n\|_2$ is $\|\Sigma^{1/2}\|_2$-Lipschitz in Z, so by Gaussian concentration (Pisier, 1989),

$$\Pr\left[\|\hat{A}_n\|_2 > \mathbb{E}\big[\|\hat{A}_n\|_2\big] + \sqrt{2\|\Sigma\|_2 t}\right] \le e^{-t}.$$

The expectation can be bounded using a result of Gordon (1985, 1988):

$$\mathbb{E}\big[\|\hat{A}_n\|_2\big] = \mathbb{E}\big[\|\Sigma^{1/2}Z\|_2\big] \le \|\Sigma^{1/2}\|_2\sqrt{n} + \|\Sigma^{1/2}\|_F.$$

Putting these together, we obtain

$$\Pr\left[\big\|\hat{\Sigma}_n\big\|_2 > \big\|\Sigma\big\|_2 + 2\sqrt{\frac{\|\Sigma\|_2\,\mathrm{tr}(\Sigma)}{n}} + 2\sqrt{\frac{2\|\Sigma\|_2^2 t}{n}} + \frac{\mathrm{tr}(\Sigma) + 2\sqrt{2\,\mathrm{tr}(\Sigma)\|\Sigma\|_2 t} + 2\|\Sigma\|_2 t}{n}\right] \le e^{-t}.$$

In our stylized scenario, this roughly implies a bound on $\|\hat{\Sigma}_n\|_2$ of the form

$$\|\Sigma\|_2 \left(1 + \sqrt{\frac{d + \mathrm{tr}(\Sigma_{d,1})/\|\Sigma\|_2}{n}} + \frac{d + \mathrm{tr}(\Sigma_{d,1})/\|\Sigma\|_2}{n}\right) + \|\Sigma\|_2\left(\sqrt{\frac{t}{n}} + \frac{t}{n}\right) \tag{5}$$

Compared to (3), we see that the main difference is that t does not scale with $\mathrm{tr}(\Sigma_{d,1})/\|\Sigma\|_2$ in (5), but it does in (3). Therefore the bounds are comparable (up to constant and logarithmic factors) when the eigenspectrum of Σ is rapidly decaying after the first d eigenvalues.

3.4.3 Approximate matrix multiplication

Finally, we give an example about approximating a matrix product $AB^\top$ using non-uniform sampling of the columns of A and B.

Example 4. Let $A := [a_1|\cdots|a_m]$ and $B := [b_1|\cdots|b_m]$ be fixed matrices, each with m columns. Assume $a_i \ne 0$ and $b_i \ne 0$ for all $i = 1,\ldots,m$. If m is very large, then the straightforward computation of the product $AB^\top$ can be prohibitive. An alternative is to take a small (non-uniform) random sample of the columns of A and B, say $a_{i_1}, b_{i_1}, \ldots, a_{i_n}, b_{i_n}$, and then compute a weighted sum of outer products

$$\frac{1}{n}\sum_{j=1}^{n}\frac{a_{i_j}b_{i_j}^\top}{p_{i_j}}$$

where $p_{i_j} > 0$ is the *a priori* probability of choosing the column index $i_j \in \{1, \ldots, m\}$ (the actual values of the probabilities p_i for $i = 1, \ldots, m$ are given below). An analysis of this scheme was given by Magen and Zouzias (2011) with the stronger requirement that the number of columns sampled be polynomially related to the allowed failure probability. Here we give an analysis in which the number of columns sampled depends only logarithmically on the failure probability.

Let $X_1, \ldots, X_n$ be i.i.d. random matrices with the discrete distribution given by

$$\Pr\left[X_j = \frac{1}{p_i}\begin{bmatrix} 0 & a_i b_i^\top \\ b_i a_i^\top & 0 \end{bmatrix}\right] = p_i \propto \|a_i\|_2 \|b_i\|_2$$

for all $i = 1, \ldots, m$, where $p_i := \|a_i\|_2\|b_i\|_2 / Z$ and $Z := \sum_{i=1}^m \|a_i\|_2\|b_i\|_2$. Let

$$\hat{M}_n := \frac{1}{n}\sum_{j=1}^n X_j \quad \text{and} \quad M := \begin{bmatrix} 0 & AB^\top \\ BA^\top & 0 \end{bmatrix}.$$

Note that $\|\hat{M}_n - M\|_2$ is the spectral norm error of approximating $AB^\top$ using the average of n outer products $\sum_{j=1}^n a_{i_j} b_{i_j}^\top / p_{i_j}$, where the indices are such that $i_j = i \Leftrightarrow X_j = a_i b_i^\top / p_i$ for $j = 1, \ldots, n$.

We have the following identities:

$$\begin{aligned}
\mathbb{E}[X_j] &= \sum_{i=1}^m p_i \left(\frac{1}{p_i}\begin{bmatrix} 0 & a_i b_i^\top \\ b_i a_i^\top & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & \sum_{i=1}^m a_i b_i^\top \\ \sum_{i=1}^m b_i a_i^\top & 0 \end{bmatrix} = M \\
\operatorname{tr}(\mathbb{E}[X_j^2]) &= \operatorname{tr}\left(\sum_{i=1}^m p_i \left(\frac{1}{p_i^2}\begin{bmatrix} a_i b_i^\top b_i a_i^\top & 0 \\ 0 & b_i a_i^\top a_i b_i^\top \end{bmatrix}\right)\right) = \sum_{i=1}^m \frac{2\|a_i\|_2^2\|b_i\|_2^2}{p_i} = 2Z^2 \\
\operatorname{tr}(\mathbb{E}[X_j]^2) &= \operatorname{tr}\left(\begin{bmatrix} AB^\top BA^\top & 0 \\ 0 & BA^\top AB^\top \end{bmatrix}\right) = 2\operatorname{tr}(A^\top AB^\top B);
\end{aligned}$$

and the following inequalities:

$$\begin{aligned}
\|X_j\|_2 &\le \max_{i=1,\ldots,m} \frac{1}{p_i}\left\|\begin{bmatrix} 0 & a_i b_i^\top \\ b_i a_i^\top & 0 \end{bmatrix}\right\|_2 = \max_{i=1,\ldots,m} \frac{\|a_i b_i^\top\|_2}{p_i} = Z \\
\|\mathbb{E}[X_j]\|_2 &= \|AB^\top\|_2 \le \|A\|_2\|B\|_2 \\
\|\mathbb{E}[X_j^2]\|_2 &\le \|A\|_2\|B\|_2 Z.
\end{aligned}$$

This means $\|X_j - M\|_2 \le Z + \|A\|_2\|B\|_2$ and $\|\mathbb{E}[(X_j - M)^2]\|_2 \le \|\mathbb{E}[X_j^2] - M^2\|_2 \le \|A\|_2\|B\|_2(Z + \|A\|_2\|B\|_2)$, so Theorem 4 and a union bound imply

$$\Pr\left[\left\|\hat{M}_n - M\right\|_2 > \sqrt{\frac{2\left(\|A\|_2\|B\|_2(Z + \|A\|_2\|B\|_2)\right)t}{n}} + \frac{(Z + \|A\|_2\|B\|_2)t}{3n}\right] \\
\le 4\left(\frac{Z^2 - \operatorname{tr}(A^\top AB^\top B)}{\|A\|_2\|B\|_2(Z + \|A\|_2\|B\|_2)}\right) \cdot t(e^t - t - 1)^{-1}.$$

Let $r_A := \|A\|_F^2 / \|A\|_2^2 \in [1, \operatorname{rank}(A)]$ and $r_B := \|B\|_F^2 / \|B\|_2^2 \in [1, \operatorname{rank}(B)]$ be the numerical (or stable) rank of A and B, respectively. Since $Z/(\|A\|_2\|B\|_2) \le \|A\|_F\|B\|_F/(\|A\|_2\|B\|_2) = \sqrt{r_A r_B}$, we have the simplified (but slightly looser) bound

$$\Pr\left[\frac{\left\|\hat{M}_n - M\right\|_2}{\|A\|_2\|B\|_2} > 2\sqrt{\frac{(1 + \sqrt{r_A r_B})(\log(4\sqrt{r_A r_B}) + t)}{n}} + \frac{2(1 + \sqrt{r_A r_B})(\log(4\sqrt{r_A r_B}) + t)}{3n}\right] \le e^{-t}.$$

Therefore, for any $\epsilon \in (0,1)$ and $\delta \in (0,1)$, if

$$n \geq \left(\frac{8}{3} + 2\sqrt{\frac{5}{3}}\right) \frac{(1 + \sqrt{r_A r_B})(\log(4\sqrt{r_A r_B}) + \log(1/\delta))}{\epsilon^2},$$

then with probability at least $1 - \delta$ over the random choice of column indices $i_1, \ldots, i_n$,

$$\left\| \frac{1}{n} \sum_{j=1}^{n} \frac{a_{i_j} b_{i_j}^\top}{p_{i_j}} - AB^\top \right\|_2 \leq \epsilon \|A\|_2 \|B\|_2. \qquad \square$$

Acknowledgements

We are grateful to Alex Gittens for useful comments and pointing out a subtle mistake in our proof of Theorem 2 in an earlier draft, and to Joel Tropp for his many comments and suggestions.

References

R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

D. A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

K. Fukumizu, F. Bach, and A. Gretton. Consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.

S. Golden. Lower bounds for the Helmholtz function. *Physical Review*, 137(4B):1127–1128, 1965.

Y. Gordon. Some inequalities for Gaussian processes and applications. *Israel J. Math.*, 50:265–289, 1985.

Y. Gordon. Gaussian processes and almost spherical sections of convex bodies. *Annals of Probability*, 16:180–188, 1988.

D. Gross. Recovering low-rank matrices from few coefficients in any basis, 2009. arXiv:0910.1879.

D. Gross, Y.-K. Liu, S. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Physical Review Letters*, 105(15):150401, 2010.

A. Guionnet. Large deviations and stochastic calculus for large random matrices. *Probability Surveys*, 1:72–172, 2004.

E. H. Lieb. Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Adv. Math.*, 11: 267–288, 1973.

A. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann. Smallest singular values of random matrices and geometry of random polytopes. *Advances in Mathematics*, 195:491–523, 2005.

A. Magen and A. Zouzias. Low rank matrix-valued Chernoff bounds and approximate matrix multiplication. In *Proceedings of the 22nd ACM-SIAM Symposium on Discrete Algorithms*, 2011.

R. I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Elec. Comm. Probab.*, 15:203–212, 2010a.

R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges, 2010b. arXiv:0911.0600.

G. Pisier. *The volume of convex bodies and Banach space geometry.* Cambridge University Press, 1989.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, 2006.

B. Recht. A simple approach to matrix completion, 2009. arXiv:0910.0651v2.

M. Rudelson. Random vectors in isotropic position. *Journal of Functional Analysis*, 164:60–72, 1999.

M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 54(4), 2007.

B. Schölkopf, A. J. Smola, and K.-R. Müller. Kernel principal component analysis. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods—Support Vector Learning*, pages 327–352. MIT Press, 1999.

C. J. Thompson. Inequality with applications in statistical mechanics. *Journal of Mathematical Physics*, 6(11):1812–1813, 1965.

J. Tropp. User-friendly tail bounds for sums of random matrices, 2011a. arXiv:1004.4389v6.

J. Tropp. Freedman's inequality for matrix martingales, 2011b. arXiv:1101.3039.

D. Voiculescu. Limit laws for random matrices and free products. *Invent. Math.*, 104:201–220, 1991.

T. Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *Proceedings of the 18th Annual Conference on Learning Theory*, 2005.

L. Zwald and G. Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In *Advances in Neural Information Processing Systems 18*. 2006.

A Sums of random vector outer products

The following lemma is a tail inequality for smallest and largest eigenvalues of the empirical covariance matrix of subgaussian random vectors. This result (with non-explicit constants) was originally obtained by Litvak et al. (2005).

Lemma 2. *Let $x_1, \ldots, x_n$ be random vectors in $\mathbb{R}^d$ such that, for some $\gamma \geq 0$,*

$$\mathbb{E}\left[x_i x_i^\top \;\middle|\; x_1, \ldots, x_{i-1}\right] = I \quad \textit{and}$$
$$\mathbb{E}\left[\exp\left(\alpha^\top x_i\right) \;\middle|\; x_1, \ldots, x_{i-1}\right] \leq \exp\left(\|\alpha\|_2^2 \gamma/2\right) \quad \textit{for all } \alpha \in \mathbb{R}^d$$

for all $i = 1, \ldots, n$, almost surely. For all $\epsilon_0 \in (0, 1/2)$ and $\delta \in (0, 1)$,

$$\Pr\left[\ \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right) > 1 + \frac{1}{1-2\epsilon_0} \cdot \epsilon_{\epsilon_0,\delta,n} \quad \textit{or} \quad \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right) < 1 - \frac{1}{1-2\epsilon_0} \cdot \epsilon_{\epsilon_0,\delta,n}\ \right] \leq \delta$$

where

$$\epsilon_{\epsilon_0,\delta,n} := \gamma \cdot \left(\sqrt{\frac{32\left(d\log(1+2/\epsilon_0) + \log(2/\delta)\right)}{n}} + \frac{2\left(d\log(1+2/\epsilon_0) + \log(2/\delta)\right)}{n}\right).$$

Remark 1. In our applications of this lemma, we will simply choose $\epsilon_0 := 1/4$ for concreteness. □

We give the proof of Lemma 2 for completeness.

The subgaussian property most readily lends itself to bounds on linear combinations of subgaussian random variables. However, we are interested in bounding certain quadratic combinations. Therefore we bootstrap from the bound for linear combinations to bound the moment generating function of the quadratic combinations; from there, we can obtain the desired tail inequality.

The following lemma relates the moment generating function to a tail inequality.

Lemma 3. *Let W be a non-negative random variable. For any $\eta \in \mathbb{R}$,*

$$\mathbb{E}\left[\exp(\eta W)\right] - \eta\mathbb{E}[W] - 1 = \eta\int_0^\infty \left(\exp(\eta t) - 1\right) \cdot \Pr[W > t] \cdot dt.$$

Proof. Integration-by-parts. □

The next lemma gives a tail inequality for any particular Rayleigh quotient of the empirical covariance matrix.

Lemma 4. *Let $x_1, \ldots, x_n$ be random vectors in $\mathbb{R}^d$ such that, for some $\gamma \geq 0$,*

$$\mathbb{E}\left[x_i x_i^\top \;\middle|\; x_1, \ldots, x_{i-1}\right] = I \quad \textit{and}$$
$$\mathbb{E}\left[\exp\left(\alpha^\top x_i\right) \;\middle|\; x_1, \ldots, x_{i-1}\right] \leq \exp\left(\|\alpha\|_2^2 \gamma/2\right) \quad \textit{for all } \alpha \in \mathbb{R}^d$$

for all $i = 1, \ldots, n$, almost surely. For all $\alpha \in \mathbb{R}^d$ such that $\|\alpha\|_2 = 1$, and all $\delta \in (0, 1)$,

$$\Pr\left[\alpha^\top\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\alpha > 1 + \sqrt{\frac{32\gamma^2\log(1/\delta)}{n}} + \frac{2\gamma\log(1/\delta)}{n}\right] \leq \delta$$

and

$$\Pr\left[\alpha^\top\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\alpha < 1 - \sqrt{\frac{32\gamma^2\log(1/\delta)}{n}}\right] \leq \delta.$$

Proof. Fix $\alpha \in \mathbb{R}^d$ with $\|\alpha\|_2 = 1$. For $i = 1, \ldots, n$, let $W_i := (\alpha^\top x_i)^2$, so $\mathbb{E}[W_i] = 1$. For any $t \geq 0$, using Chernoff's bounding method gives

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left[W_i > t\right] \mid x_1, \ldots, x_{i-1}\right] \\
&\leq \inf_{\eta > 0} \left\{ \mathbb{E}\left[\mathbb{1}\left[\exp\left(\eta |\alpha^\top x_i| \right) > e^{\eta\sqrt{t}} \right] \;\middle|\; x_1, \ldots, x_{i-1} \right] \right\} \\
&\leq \inf_{\eta > 0} \left\{ e^{-\eta\sqrt{t}} \cdot \left(\mathbb{E}\left[\exp\left(\eta \alpha^\top x_i \right) \;\middle|\; x_1, \ldots, x_{i-1} \right] + \mathbb{E}\left[\exp\left(-\eta \alpha^\top x_i \right) \;\middle|\; x_1, \ldots, x_{i-1} \right] \right) \right\} \\
&\leq \inf_{\eta > 0} \left\{ 2\exp\left(-\eta\sqrt{t} + \eta^2\gamma/2 \right) \right\} \\
&= 2\exp\left(-\frac{t}{2\gamma} \right).
\end{aligned}$$

So by Lemma 3, for any $\eta < 1/(2\gamma)$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\eta W_i\right) \mid x_1, \ldots, x_{i-1}\right] &\leq 1 + \eta + \eta \int_0^\infty \left(\exp\left(\eta t\right) - 1\right) \cdot 2\exp\left(-\frac{t}{2\gamma} \right) \cdot dt \\
&= 1 + \eta + \frac{8\eta^2\gamma^2}{1 - 2\eta\gamma} \\
&\leq \exp\left(\eta + \frac{8\eta^2\gamma^2}{1 - 2\eta\gamma} \right)
\end{aligned}$$

and therefore

$$\mathbb{E}\left[\exp\left(\eta \sum_{i=1}^n W_i \right) \right] \leq \exp\left(n\eta + \frac{8n\eta^2\gamma^2}{1 - 2\eta\gamma} \right).$$

Using Chernoff's bounding method twice more gives

$$\begin{aligned}
\Pr\left[\sum_{i=1}^n W_i > n + t \right] &\leq \inf_{0 \leq \eta < 1/(2\gamma)} \left\{ \exp\left(-t\eta + \frac{8n\eta^2\gamma^2}{1 - 2\eta\gamma} \right) \right\} \\
&= \exp\left(-\frac{8n\gamma^2 + \gamma t - \sqrt{8n\gamma^2\left(8n\gamma^2 + 2\gamma t\right)}}{2\gamma^2} \right)
\end{aligned}$$

and

$$\Pr\left[\sum_{i=1}^n W_i < n - t \right] \leq \inf_{\eta \leq 0} \left\{ \exp\left(t\eta + \frac{8n\eta^2\gamma^2}{1 - 2\eta\gamma} \right) \right\} \leq \exp\left(-\frac{t^2}{32n\gamma^2} \right).$$

The claim follows. $\square$

In order to bound the smallest and largest eigenvalues of the empirical covariance matrix, we apply the bound for the Rayleigh quotient in Lemma 4 together with a covering argument.

Lemma 5 (Pisier, 1989)**.** *For any $\epsilon_0 > 0$, there exists $Q \subseteq \mathcal{S}^{d-1} := \{\alpha \in \mathbb{R}^d : \|\alpha\|_2 = 1\}$ of cardinality $\leq (1 + 2/\epsilon_0)^d$ such that $\forall \alpha \in \mathcal{S}^{d-1} \exists q \in Q \,\bullet\, \|\alpha - q\|_2 \leq \epsilon_0$.*

Proof of Lemma 2. Let $\hat{\Sigma} := (1/n)\sum_{i=1}^n x_i x_i^\top$, let $\mathcal{S}^{d-1} := \{\alpha \in \mathbb{R}^d : \|\alpha\|_2 = 1\}$ be the unit sphere in $\mathbb{R}^d$, and let $Q \subset \mathcal{S}^{d-1}$ be an ϵ_0-cover of $\mathcal{S}^{d-1}$ of minimum size with respect to $\|\cdot\|_2$. By Lemma 5, the cardinality of Q is at most $(1+2/\epsilon_0)^d$. Let E be the event

$$\max\left\{|q^\top(\hat{\Sigma}-I)q| : q \in Q\right\} \le \epsilon_{\epsilon_0,\delta,n}.$$

By Lemma 4 and a union bound, $\Pr[E] \ge 1-\delta$. Now assume the event E holds. Let $\alpha_0 \in \mathcal{S}^{d-1}$ be such that $|\alpha_0^\top(\hat{\Sigma}-I)\alpha_0| = \max\{|\alpha^\top(\hat{\Sigma}-I)\alpha| : \alpha \in \mathcal{S}^{d-1}\} = \|\hat{\Sigma}-I\|_2$. Using the triangle and Cauchy-Schwarz inequalities, we have

$$\begin{aligned}
\|\hat{\Sigma}-I\|_2 = |\alpha_0^\top(\hat{\Sigma}-I)\alpha_0| &= \min_{q\in Q} |q^\top(\hat{\Sigma}-I)q + \alpha_0^\top(\hat{\Sigma}-I)\alpha_0 - q^\top(\hat{\Sigma}-I)q| \\
&\le \min_{q\in Q} |q^\top(\hat{\Sigma}-I)q| + |\alpha_0^\top(\hat{\Sigma}-I)\alpha_0 - q^\top(\hat{\Sigma}-I)q| \\
&= \min_{q\in Q} |q^\top(\hat{\Sigma}-I)q| + |\alpha_0^\top(\hat{\Sigma}-I)(\alpha_0-q) - (q-\alpha_0)^\top(\hat{\Sigma}-I)q| \\
&\le \min_{q\in Q} |q^\top(\hat{\Sigma}-I)q| + \|\alpha_0\|_2\|\hat{\Sigma}-I\|_2\|\alpha_0-q\|_2 + \|q-\alpha_0\|_2\|\hat{\Sigma}-I\|_2\|q\|_2 \\
&\le \epsilon_{\epsilon_0,\delta,n} + 2\epsilon_0\|\hat{\Sigma}-I\|_2
\end{aligned}$$

so

$$\max\left\{\lambda_{\max}(\hat{\Sigma})-1,\ 1-\lambda_{\min}(\hat{\Sigma})\right\} = \|\hat{\Sigma}-I\|_2 \le \frac{1}{1-2\epsilon_0}\cdot\epsilon_{\epsilon_0,\delta,n}.$$

□